基于 LPCA 的谱聚类算法 *

童　涛，文国秋[†]，谭马龙，吴　林，杜婷婷

(广西师范大学 广西多源信息挖掘与安全重点实验室, 广西 桂林 541004)

摘　要：针对传统谱聚类在构建关系矩阵时只考虑样本的全局特征而忽略样本的局部特征、在聚类划分时通常需要指定聚类个数、无法对交叉点进行正确划分等问题，提出了一种改进的基于局部主成分分析和连通图分解的谱聚类算法。首先自动学习挑选数据集的中心点，然后使用局部主成分分析得到数据集的关系矩阵，最后用连通图分解算法完成对关系矩阵的划分。实验结果表明提出的改进算法性能优于现有经典算法。

关键词：局部主成分分析；谱聚类；连通图分解；交叉点

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2018.04.0283

Spectral clustering algorithm based on LPCA

Tong Tao, Wen Guoqiu[†], Tan Malong, Wu Lin, Du Tingting

(*Guangxi Key Laboratory of Multi-source Information Mining & Security Guangxi Normal University*, *Guilin Guangxi* 541004, *China*)

Abstract: As the traditional spectral clustering algorithms 1) only considered the global structures of the samples while ignoring their local structures for the construction of the correlation matrix; 2) conducted clustering with a predefined cluster number; 3) could not divide the intersections correctly. This paper proposes a new method based on the local principal component analysis and the decomposition method of the connected graph. Specifically, the proposed method automatically learns the centroids of the selected subset of the samples, obtains the correlation matrix of the samples based on the local principal component analysis, and uses the decomposition method of the connected graph to partition the resulting correlation matrix. Experimental results show that the proposed algorithm performs better than the existing algorithms.

Key words: local principle content analysis; spectral clustering; connected graph decomposition; intersection

0　引言

聚类[1~6]是将相同或者相似的样本划分到同一类或簇，不同样本划分到不同的类或簇的一种常见数据处理技术。根据聚类方式，现有聚类算法可以分为划分式聚类、层次化聚类、基于密度和网格的聚类等类别[2]。其中，K-means 思路简单且易于实现，已经得到了广泛的应用。但是 K-means 有两个较为突出的问题，即聚类中心的初始化和聚类个数的确定。常见的 K-means 方法采用各式各样的聚类中心初始化方法，不同的初始化导致不同的聚类结果。另外，人工指定类数的方法需要经验或者对数据分布具有先验知识。

为了解决 K-means 自身存在的问题，研究者提出了各种改进的 K-means 算法。例如，针对 K-means 无法对非凸数据集进行聚类的问题[4]，通过引入高斯分布函数[5]使得该算法突破了原始数据集分布的限制，但它仍然需要指定聚类个数 k。为了同时解决 K-means 算法的两个问题[7]，提出了通过样本密度来确定中心点，将中心点附近的样本划分为该中心点类别的算法。此方法使传统 K-means 的两大问题都得到了解决，但这种方法需要计算所有样本的密度和距离，需耗费大量的时间且该方法也没有解决交叉点划分困难的问题。

本文提出了一种基于局部主成分分析和连通图分解的谱聚类算法（spectral clustering based on local principle content analysis,SC-LPCA）。具体地说，首先随机选出数据集中一部分数据作为新的数据集，再对新数据集中的每个样本求出其邻域样本构成的矩阵，然后对该矩阵集合进行 LPCA（local principle component analysis）处理得到数据集的关系矩阵，接着使用连

收稿日期：2018-04-17；**修回日期：**2018-05-28　**基金项目：**国家重点研发计划资助项目（2016YFB1000905）；国家自然科学基金资助项目（61170131，61263035，61573270，90718020）；国家"973"计划资助项目（2013CB329404）；中国博士后科学基金资助项目（2015M570837）；广西自然科学基金资助项目（2015GXNSFCB139011，2015GXNSFAA139306）

作者简介：童涛（1990-），男，湖北仙桃人，硕士，主要研究方向为机器学习和数据挖掘；文国秋（1987-），女（通信作者），硕士，主要研究方向为机器学习和数据挖掘（1326261060@qq.com）；谭马龙（1993-），男，湖北襄阳人，硕士，主要研究方向为机器学习和数据挖掘；吴林（1993-），女，安徽安庆人，硕士，主要研究方向为机器学习和数据挖掘；杜婷婷（1993-），女，山西大同人，硕士，主要研究方向为机器学习和数据挖掘.

通图分解算法对关系矩阵进行划分，最后以这些选取的点为中心依照距离来划分剩余所有点，得到最后聚类结果。

本文算法相较于传统的聚类算法的优势在于：a）通过挑选聚类的中心点而不是直接聚类整个数据集，大幅度地减少了聚类的计算量；b）通过使用 LPCA 使得到的关系矩阵较好地描述了数据集的局部特征，提升了算法聚类性能；c)利用数据挑选和 LPCA 处理,解决了多流形数据集的交叉点划分困难甚至错误的问题；d)使用连通图分解算法，不用指定聚类个数，就可以完成聚类，降低了聚类复杂性和难度。

1 相关理论

1.1 局部主成分分析

主成分分析[1,8] (principal component analysis, PCA) 是一种将高维数据投影到低维数据空间的方法。给定一个样本数据集 $X=[x_1,x_2,...,x_d]\in\mathbb{R}^{n\times d}$ （d 是样本的属性数，n 是样本个数），对样本集先进行中心化（$\Sigma_i x_i=0$）然后求出数据集的协方差再对它进行投影。对得到的投影后的新的属性进行排序，取值大的前 d' （$d'<d$）个属性。这样就得到了样本点 x_i 在低维坐标系中的投影 $z_i=(z_{i1},z_{i2},...,z_{id'})$，其中 $z_{ij}=w_j^T x_i$，是样本 x_i 在低维坐标系下第 j 维的值。由于协方差可以衡量样本间的离散程度，而 PCA 利用了样本的全局的协方差。因此得到的结果较好的保留了原始数据集的全局结构特征。

局部主成分分析[9]（LPCA）它是在传统主成分分析上的一种改进。其分析对象不再是整个数据集的分布情况，而是研究单个数据与其周围邻域数据之间的分布情况。通过对样本邻域所构成的子集去做协方差处理和特征值分解，使得到的结果能更好的反映样本与其周围样本之间的关系。通过 LPCA 处理得到的关系矩阵能更好的反映数据集的真实结构，也能更好的体现样本间的局部特征。

1.2 谱聚类

谱聚类[10-13]（spectral clustering, SC）是一种利用图论的思想，把聚类转换成了图的分割问题的聚类方法。给定图 G，把数据集 $X\in\mathbb{R}^{n\times d}$ 中每一个样本当作一个点 V，样本点之间的相关性定义为边 E，这样就形成了该数据集合构成的图 $G(V,E)$。然后依据此图构造关系矩阵 W，通常用欧式距离来描述样本之间的相关性，即 $w_{ij}=\|x_i-x_j\|_2$。我们把一个点的所有与之相连的边的权值相加得到的结果称为该点的度记为 de_i，表示如下：

$$de_i=\sum_{j=1}^{n} w_{ij} \tag{1}$$

把所有样本点的度构成的矩阵称为度矩阵 D。利用关系矩阵 W 和度矩阵 D 计算得到拉普拉斯矩阵 $L=D-W$。然后对 L 求协方差并进行特征值分解，取得到的结果的前 d' 个特征值所对应的特征向量构成新的特征矩阵 F，最后把 F 的每一行当成一个新的样本 $\hat{x}_i\in\mathbb{R}^{1\times d'}$，使用 K-means 对 F 进行划分，即为最终的聚类结果。谱聚类最大的特点是它通过谱图的引入巧妙地解决了以前直接使用 K-means 聚类时存在的无法处理非凸数据集的问题。

基于 LPCA 的谱聚类算法，是在传统谱聚类的基础之上通过引入数据样本在局部的分布特性，来对数据集进行分析。这样可以比较充分地利用数据在局部所具有的特性，而不是直接以数据集为一个整体去分析数据集的分布特点以及数据样本之间的关系，这样能更加充分地利用原始数据所包含的信息。得到的关系矩阵更能表征原始的数据集，这对于提升算法的聚类性能有很大的帮助。

2 算法描述

本文提出的 SC-LPCA 算法通过引入中心样本的邻域子集和 LPCA 处理，保留了数据集的局部特征，并且通过连通图的分解使得聚类不需要指定类的个数就可以自动完成。下文将详细介绍 SC-LPCA 算法步骤。

给定样本集 $X=[x_1,x_2,...,x_d]\in\mathbb{R}^{n\times d}$，随机挑选一个样本记为 y_i，邻域阈值记为 r 表达式如下：

$$r=\text{mean}\,\|x_i-x_j\|_2\ (i,j\in[1,n]) \tag{2}$$

样本 y_i 的邻域子集记为 $Nr(y_i)$ 表达式如下：

$$N(x)=\{x_j:\|x-x_j\|\le r\} \tag{3}$$

以 y_i 为中心，随机挑选一个不在 $Nr(y_i)$ 中的样本记为 y_{i+1} 即：$y_{i+1}\notin Nr(y_i)$。重复该挑选操作 n_0 次，得到原始数据集的中心点构成的新数据集 $Y=[y_1,y_2,...,y_{n_0}]\in\mathbb{R}^{n_0\times d}$。对得到的新数据集 Y 的每个样本的邻域所构成的矩阵，求其协方差记为 C_i，表达式如下：

$$C_i=y_i^T\bullet y_i \tag{4}$$

对 C_i 进行特征值分解，取分解后值较大的前 d'（$d'\in[1,d-1]$）个特征值对应的特征向量所构成的矩阵记为 Q_i。样本的空间阈值记为 ε 表达式如下：

$$\varepsilon=\max_{1\le i\le n_0}\min_{i\ne j}\|y_i-y_j\| \tag{5}$$

样本的投影规模阈值记为 η，表达式如下：

$$\eta=\underset{(i,j):\|y_i-y_j\|<\varepsilon}{\text{median}}\|Q_i-Q_j\| \tag{6}$$

依据得到的数据集 Y 和投影结果 Q 计算出关系矩阵 W，表达式如下：

$$w_{ij}=\exp(-\frac{\|y_i-y_j\|^2}{\varepsilon^2}).\exp(-\frac{\|Q_i-Q_j\|^2}{\eta^2}) \tag{7}$$

把用于对 W 进行 0,1 化的阈值记为 δ，其表达式如下：

$$\delta=NUM\times(\text{median}\ W)\ (NUM\in(0,1)) \tag{8}$$

对 W 进行划分的表达式如下：

$$w_{i,j}^{*}=\begin{cases}1, \text{if} \quad 0<w_{i,j}<\delta \\ 0, \text{otherwise}\end{cases} \tag{9}$$

使用得到的关系矩阵 W^* 构造出连通图集合，对每个连通图进行递归分解，对得到的最大连通图计算其分裂阈值 λ 与承受阈值 t 的关系来决定是否分解。t 的表达式如下：

$$t=\min\{u_1,u_2\}/n_u \tag{10}$$

其中：u_1 与 u_2 表示分割后的两部分的点的个数，而 n_u 表示两个连接部分的边的数量。λ 的表达式如下：

$$\lambda=\frac{1}{2}e^{\frac{a}{b}} \tag{11}$$

其中：a 表示最大连通图中边的数量，b 表示最大连通图中点的数量。依据 λ 与 t 的关系分解完所有连通图集合，即所有中心点得到正确划分。最后按照离中心点的距离划分剩余所有的点的类别。

算法流程如下：

SC-LPCA 算法伪代码

输入：训练集 $X\in\mathbb{R}^{n\times d}$，特征值向量维数 d'。

输出：聚类结果。

1. 随机挑选 y_1 然后再随机挑选不在 $Nr(y_1)$ 中的样本为 y_2，重复 n_0 次得到挑选后的数据集 $Y\in\mathbb{R}^{n_0\times d}$；
2. 对每一个样本 y_i 计算它的邻域子集 $Nr(y_i)$ 的协方差值 C_i，对 C_i 进行特征值分解取最大的 d' 个特征值对应的特征向量所构成的矩阵记为 Q_i；
3. 按照公式（7）计算 W；
4. 按照公式（9）用 δ 对 W 进行 0, 1 化，得到新的关系矩阵 W^*；
5. 使用连通图分解算法对 W^* 进行聚类划分；
6. 以得到的划分结果为中心点计算剩余所有样本到中心点的距离，样本所属类别就是离它最近中心点的类别。

本文所提出的算法选择的不是直接对原始数据集 X 进行相关的聚类操作，而是先通过挑选得到新的数据集 Y，再对 Y 进行聚类。首先对挑选出来的中心点聚类然后以中心点为基础，依照就近原则划分剩余样本，即剩余样本所属的类别就是离它最近中心点的类别。这样不仅缩小了聚类的规模降低了计算量，而且还可以让聚类的中心点不至于过分集中使聚类的划分更精确。如此，既提高了算法的效率又提高了算法的聚类准确性。

通过对挑选后的数据集样本的邻域所构成的矩阵集合进行 LPCA 处理，使得到的关系矩阵更好的保留了原始数据集样本的局部特征。除此之外投影也降低了待计算数据集的规模，进一步降低了计算量，这在处理高维数据时可以减少计算所需的时间。

通过使用中心点挑选配合 LPCA 处理，本文巧妙地解决了真实数据集中经常会出现的交叉点聚类困难甚至错误的问题。首先，通过挑选随机点邻域外的点为中心点使得异类样本尽量分开，即不同的类之间关系变得较为松散。其次，通过对挑选得到的样本的邻域所构成的矩阵做协方差处理，使得局部样本尽量聚拢，即同类样本内部的关系变得更加紧密。如此，虽然能使交叉点彼此达到一定程度上的分离，但是当不同簇的夹角较小的时候单独使用协方差不一定能很稳定的实现交叉点的分离。如图 1 所示，当两个簇的夹角 θ 足够小的时候样本 p_1 与 p_2 之间的距离就可能比样本 p_1 与 p_0 之间的距离小，这样在以距离来划分样本类别的时候可能就出现错误划分的情况。因此在引入协方差的同时，为了稳定对交叉点的划分效果同时引入了投影。通过投影可以使同类样本投影后更加亲密，异类样本投影后更加疏远。这样就较好的解决了可能出现的不同的簇夹角过小带来的交叉点划分困难的问题，因此本文引入了 LPCA 配合样本挑选来解决交叉点聚类困难的问题。

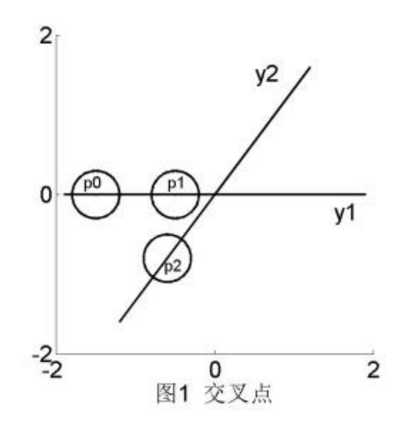


图1 交叉点

在对挑选后的数据集进行 LPCA 处理后，得到了能反映原始数据集的真实分布情况的关系矩阵。由于初始本文用来描述样本之间的关系使用的是欧式距离所形成的矩阵，因此在对这个关系矩阵不断的操作后，得到的关系矩阵仍然是一个实数形成的关系矩阵。然而在使用这个关系矩阵构建数据集的连通图的时候所有数据之间的距离一定都不为 0，即所有样本彼此都是有关系的。因此，本文使用了阈值 δ 对这个实数形成的关系矩阵按照公式（9）进行处理，就得到了能较好描述数据集样本间关系的 0、1 矩阵（0 表示两个样本之间无关系，1 表示两个样本之间有关系）。这样就使得原始的全连通图得到了一定程度上的分解，即通过这个 0、1 矩阵得到了能表征原始数据集的结构特征的连通图集合。

最后对于得到的连通图集合中的每一个连通图，递归寻找其最大子连通图。找到最大子连通图后，计算在此情形下想要进行分裂的部分与剩余部分所计算出的分裂阈值 λ 与承受阈值 t 是否满足关系式：$\lambda>t$，当该式子成立时就将该连通图分解成两个部分。对分解完成后，剩余的部分同样进行递归分解，一直到整个连通图所有可能的分解结束。如此，对整个连通图集合完成分解，得到的每一个子连通图就对应聚类结果的一个类。这样通过寻找子连通图，直接完成了最后的聚类划分。由于省去了对数据集的真实类个数 k 的寻找问题，以及不确定是否在划分为 k 类时聚类效果最佳的问题。因此，SC-LPCA 算法不但降低了聚类的难度而且提高了聚类的准确率。

经过前面的步骤，挑选出来的中心点已经得到了正确的划

分。对于剩下来的样本数据，直接用欧氏距离来衡量样本点与中心点的距离，然后将剩余样本划入到离它最近的中心点的类别中去。如此就省去了繁复的计算，完成了对剩余样本的聚类。

3　实验结果与分析

本文提出的算法（SC-LPCA）与各个对比算法均使用 MATLAB 2014a 编程实现，且所有实验均是在 Win10，64 位操作系统下测试完成。本实验所使用的硬件环境为：CPU: Intel(R) Core™ i7-7700 CPU @ 3.6 GHz，内存:8 GB。

3.1　对比算法与评价指标

为了更好地衡量本文提出的算法的性能，将该算法与其他聚类算法进行对比，如 K-means、LSR、SSC 等，具体对比算法详细信息如下：（本文将 K-means 这个经典的聚类算法作为基准线（baseline）来衡量所有的算法的好坏）

K-means [1,4]是通过人工指定的 k 值随机挑选 k 个样本为初始中心点，通过计算各个数据点距中心点的距离来对所有数据进行聚类，然后计算本次聚类后各类的平均值为新的中心点并以此更新旧的中心点，重复迭代上述过程至聚类结果稳定即为最终聚类结果。

LRR[14-15]（low rank representation），首先对数据集的谱图所表征的关系矩阵进行低秩处理，然后调用传统聚类方法对处理后的关系矩阵进行聚类划分并输出最后结果。

LSR[16-19]（least squares regression），首先对样本数据集对应的关系矩阵用 F 范数进行约束使得到的矩阵具有更好的内聚性，然后再调用传统聚类的方法再对处理后的关系矩阵进行聚类划分。

SSQP[20]（subspace segmentation via quadratic programming），首先求出其他样本与数据样本之间的线性关系，并使用 F 范数来约束该关系，求出这个样本数据集的关系矩阵，然后调用传统聚类的方法对得到的关系矩阵进行聚类划分并输出聚类结果。

NCut[21,22]（normalized cut），通过对数据集构造的带权重的无向图的分割过程进行约束，使得到的分割后的块之间的所有样本的割的和最小。然而标准的归一化割的求解是 NP-hard 问题，是无法求解的。本文所使用的对比算法使用的是将归一化割的求解引入拉普拉斯矩阵，转换成对图的度的求解，和特征值的求解。最后使用 K-means 完成对特征向量所构成的特征矩阵的聚类划分并输出聚类结果。

SSC[23-27]（sparse subspace clustering）通过对样本数据集对应的关系矩阵进行约束使得到的关系矩阵具有子空间的稀疏性，然后再调用传统的聚类方法对处理后的数据集进行聚类划分并输出聚类结果。

3.2　实验设置与数据集

8 个实验数据集都来自 UCI[1]，数据集的详细信息如表 1 所示。本文采用数据集有不同规格，这样能全面评测本文提出算法的有效性和可靠性。

表 1　数据集规模信息

数据集名称	样本个数	属性数	类个数
Arrhythmia	452	279	13
Lungdiscrete	73	325	7
YaleB	640	2016	10
Cars	392	8	3
Breast	699	10	2
Auto	205	25	6
Balance	625	4	3
Crx	690	15	2

对比算法中的各参数依据对应算法的文献进行设定，其中所有使用 K-means 直接聚类或进行聚类划分的地方，k 值均为数据集的真实类的个数。虽然 SC-LPCA 算法在聚类过程中不需要人为指定聚类个数，但是为了与对比算法有更加直观的比较，表 1 与 2 中的结果都是在聚类个数是真实类情况下得到的。具体的说，算法中提到的含公式的参数，它们的取值大都直接按公式计算得到，少数参数会在公式计算的结果上进行一定的缩放。例如 δ 的取值按式（8）计算得到，但是在具体实验调试过程中一般针对不同的数据集，为了得到更好的聚类效果，需要对该参数进行一定程度的缩小，通常取该值的 $10^{-1}\sim10^{-10}$ 倍。同样，挑选样本邻域的阈值 r 取值通常是按式（2）计算得到，但是有时为了得到更好的聚类结果也会取该值的 0.5～1.5 倍。

其他参数如 n_0 取值使用的是 10%~70%总样本数中挑选出的较优的值。使用 LPCA 进行投影时，取特征值分解后的最大的 d' 个属性，其中 d' 的取值范围是 $d' \in [1, d-1]$ 中的较优值。本文采用的评价指标包括 ACC（accuracy，准确率），NMI（normalized mutual information，标准互信息）。其定义形式如下：

$$\mathrm{ACC} = \frac{1}{n}\sum_{i=1}^{n} 1(l_i = \hat{l}_i) \tag{12}$$

在这里 l_i 是样本的真实标签，$\hat{l}$ 表示第 i 个样本的预测标签，n 是样本个数。

$$\mathrm{NMI}(U,V) = \frac{\mathrm{MI}(U,V)}{\sqrt{H(U)H(V)}} \tag{13}$$

$\mathrm{MI}(U,V)$ 表示标签 U(真实标签)和标签 V(预测标签)之间的互信息，$H(U)$ 表示标签 U 的熵，$H(V)$ 表示标签 V 的熵。

3.3　实验结果与分析

所有算法在 8 个数据集上的计算结果如表 2、3 和图 2 所示。

表 2　不同数据集下各算法的 ACC

1　http://archive.ics.uci.edu/ml/index.php

数据集名称	K-means	LRR	LSR	SSQP	NCut	SSC	SC-LPCA
Arrhythmia	0.2235	0.2544	0.2522	0.2323	0.2876	0.2367	**0.3111**
Lungdiscrete	0.5616	0.7260	0.7260	0.8082	0.8630	0.7260	**0.8750**
YaleB	0.1891	0.3047	0.2859	0.2234	0.2891	0.2594	**0.4219**
Cars	0.4490	0.4847	0.4643	0.5383	0.5153	0.5740	**0.7092**
Breast	0.6009	0.6381	0.6381	0.6381	0.6009	0.6381	**0.6810**
Auto	0.3220	0.3415	0.3268	0.3951	0.3512	0.3463	**0.5238**
Balance	0.5120	0.5252	0.5296	0.5248	0.5168	0.5488	**0.5556**
Crx	0.5333	0.5580	0.5435	0.5493	0.5478	0.5580	**0.6087**

表 3　不同数据集下各算法的 NMI

数据集名称	K-means	LRR	LSR	SSQP	NCut	SSC	SC-LPCA
Arrhythmia	0.1990	0.2309	0.2173	0.2186	0.2322	0.2126	**0.2369**
Lungdiscrete	0.5891	0.5988	0.6276	0.7619	0.7995	0.6339	**0.9091**
YaleB	0.1160	0.2408	0.2308	0.1320	0.2385	0.2033	**0.3217**
Cars	0.2049	0.2646	0.2702	0.2126	0.2202	0.2352	**0.3453**
Breast	0.1261	**0.6169**	0.6169	0.6169	0.1261	0.6169	0.3443
Auto	0.1000	0.1450	0.0626	0.1544	0.1533	0.1504	**0.3615**
Balance	0.1009	0.1510	0.1483	0.1117	0.1019	0.1241	**0.1710**
Crx	0.4439	0.6429	0.4978	0.6065	0.5397	0.6429	**0.6483**

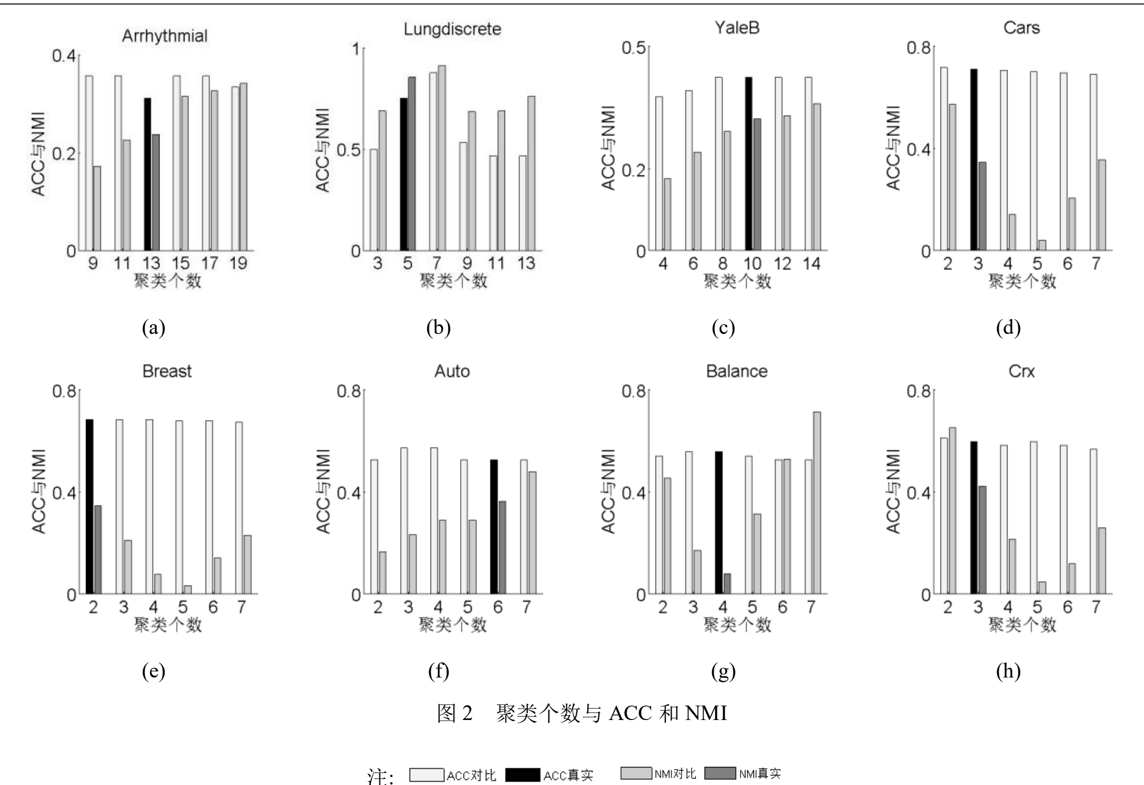


图 2　聚类个数与 ACC 和 NMI

从表 2 中可以看出本文所提出的算法在准确率上相较对比算法在 8 个数据集上都有不错的提升。例如对于数据集 Cars 而言，本文所提算法与对比算法相比较在 ACC 上的提升为 26.02%、22.45%、24.49%、17.09%、19.39%、13.52%平均提升有 20.49%。这些结果表明本文提出的算法在我们进行实验的数据集上与对比算法相比都有一定的提升，这也从侧面反映了本文所提算法的合理性和有效性。例如在 Cars 上表现最好的 SSC 算法，它通过稀疏子空间的约束，使得到的关系矩阵较传统谱聚类所使用的拉普拉斯矩阵更能表征数据集的结构特点。因此它会比传统谱聚类算法准确率更高。然而该算法只考虑了样本的全局特性而忽视了样本的局部特性因而用来描述数据集的关系矩阵就不是那么精确，这使得聚类结果并不完全准确。而本文提出的聚类算法利用 LPCA 处理，较好的描述了数据样本的局部特性，尽可能多地维持了关系矩阵信息的丰富性，即尽可

能多地保留了原始数据集的信息，使获得的聚类结果更加准确。

从表3中可以看出，本文所提出的算法除了在数据集Breast上不是最好，在其他数据集上均好于对比算法。同样以 Cars 数据集为例，本文所提出的算法与对比算法相比较在 NMI 上的提升为 14.04%、8.07%、7.51%、13.27%、12.54%、11.01%平均提升为 11.07%。这表明本文所使用的 SC-LPCA 算法得到的关系矩阵，更好的反映了数据样本之间的关系，因此它在标准互信息的值上比其他对比算法更高。同样对 Cars 数据集而言，以在 NMI 上表现最好的算法 LSR 为例。它通过约束提高了关系矩阵的内聚性，使得和其他对比算法相比较，它的 NMI 值会更高。但是它仍然是从全局出发来考虑数据集的特性，并没有考虑数据的其他同样重要的特征，比如数据集所含的交叉点分布的情况。而本文提出的算法，通过对中心点的挑选配合 LPCA 处理以及最后按距离对剩余样本的划分，不仅考虑了样本的局部特征，而且使数据集中的交叉点得到了较好的划分，因此本文提出的算法会在 NMI 上比其他对比算法更好。

从表4中可以看出，本文提出的算法除了在数据集Breast上不是最好，在其他数据集上均好于对比算法。以 Cars 数据集为例，本文所提算法与对比算法相比较在执行时间上的提升为 0.06s、105.388s、0.491s、0.428s。这表明通过数据集的挑选来先期对代表性的样本聚类，然后后期依照距离对剩余样本聚类确实在一定程度上缩短了聚类所用的时间。而且，降低聚类数据集的维度也在一定程度上减少了聚类的计算量，从而也达到了提升了算法的执行时间的目的。

表 4 不同数据集下各算法的执行时间 /s

数据集名称	LRR	SSQP	NCut	SSC	SC-LPCA
Arrhythmia	6.602	127.139	1.697	2.775	**1.152**
Lungdiscrete	0.327	0.178	0.500	0.142	**0.090**
YaleB	40.511	120.105	16.41	19.36	**14.772**
Cars	0.327	105.655	0.758	0.695	**0.267**
Breast	**0.502**	392.849	0.538	2.168	0.544
Auto	0.405	17.277	0.694	0.271	**0.087**
Balance	0.367	52.274	1.189	0.821	**0.361**
Crx	0.493	371.820	0.463	1.822	**0.387**

为了更好地反映聚类个数对聚类性能的影响，本文在实验过程中通过调节算法中的样本个数阈值 n_0、邻域阈值 r、空间阈值 ε 和投影规模阈值 η 以及 0,1 化 $\boldsymbol{W}$ 的阈值 δ 使得算法能得到不同的聚类个数。具体在得到不同类数的情况下算法性能如图 2 所示。结果表明在这 8 个数据集上，本文提出的算法并不一定都是在得到真实类的情况下聚类性能最佳。例如对于 Auto 数据集而言是在得到 4 类的情况下取得最佳聚类效果，而不是在真实类 6 聚类效果最佳。本文通过对数据集合所构成的连通图的分解而自动地对数据集进行聚类划分，不用去人为指定聚类个数，因此能获得更好的聚类结果。

4 结束语

本文是对传统谱聚类算法的改进，它先通过挑选有代表性的样本数据进行聚类，然后再推广到其他所有样本。这使得提出的算法能有效降低聚类的计算量，同时通过利用 LPCA 处理保持了数据的局部特征提高了聚类结果的准确性同时也降低了聚类的规模。此外，算法依据数据的分布情况进行聚类，而不必像 K-means 一样需要事先知道数据的真实类数，有效提高了真实应用中的实用性。联合使用中心点挑选和 LPCA 处理，解决了交叉点划分困难的问题。通过在多个数据集和多个对比算法的比较分析，提出算法能在低维数据集上获得比较好的聚类结果，同时也证明了聚类算法不一定都在真实类数下获得最好聚类结果。在未来工作中需要进一步考虑算法在处理高维大数据，甚至超高维大数据的能力，以及对含噪声的数据处理的能力。

参考文献：

[1] 周志华. 机器学习 [M]. 北京：清华大学出版社, 2016: 197-217, 229-231. (Zhou Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016: 197-217, 229-231)

[2] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究 [J]. 软件学报, 2008, 19 (1): 48-61. (Sun Jigui, Liu Jie, Zhao Lianyu. Clustering algorithms research [J]. Journal of Software, 2008, 19 (1): 48-61.)

[3] 李永钢, 苏毅娟, 何威, 等. 基于超图和样本自表征的谱聚类算法 [J]. 计算机应用研究, 2017, 34 (6): 1621-1625. (Li Yonggang, Su Yijuan, He Wei, *et al.* Hypergraph and self-representation for spectral clustering [J]. Application Research of Computers, 2017, 34 (6): 1621-1625.)

[4] 王俊杰. 密度敏感的K-means聚类算法研究 [D]. 济南：山东师范大学, 2014. (Wang Junjie. Density-sensitive K-means clustering algorithm [D]. Jinan: Sandong Normal University, 2014.)

[5] 尹楠. 基于高斯混合模型的期望最大化聚类算法 [J]. 统计与决策, 2017 (4): 87-89. (Yin Nan. Expectation maximization clustering algorithm based on Gauss mixture model [J]. Statistic and Decision, 2017 (4): 87-89.)

[6] 邓健爽, 郑启伦, 彭宏, 等. 基于连通图动态分裂的聚类算法 [J]. 华南理工大学学报：自然科学版, 2007, 35 (1): 118-122. (Deng Jianshuang, Zheng Qilun, Peng Hong, *et al.* Clusteirng algorithm based on dynamic division of connected graph [J]. Journal of South China University of Technology: Natural Science Edition, 2007, 35 (1): 118-122.)

[7] Rodriguez A, Laio A. Clustering by fast search and find of density peaks [J]. Science, 2014, 344 (6191): 1492.

[8] Mika S, Smola A, Scholz M. Kernel PCA and de-noising in feature spaces [C]// Proc of Conference on Advances in Neural Information Processing Systems II. Cambridge: MIT Press, 1999: 536-542.

[9] Arias-Castro E, Lerman G, Zhang T. Spectral clustering based on local PCA [J]. Journal of Machine Learning Research, 2017, 18 (1): 253-309.

[10] Yu X S, Shi Jianbo. Multiclass spectral clustering [C]// Proc of the 9th IEEE

International Conference on Computer Vision. 2003: 313-319.

[11] He Xin, Wang Jiabing, Zhang Zhongxian, *et al.* Clustering Web documents based on Multiclass spectral clustering [C]// Proc of International Conference on Machine Learning and Cybernetics. 2011: 1466-1471.

[12] Zhu Xiaofeng, He Wei, Li Yonggang, *et al.* One-step spectral clustering via dynamically learning affinity matrix and subspace [C]// Proc of the 31st AAAI Conference on Artificial Intelligence. 2007: 2963-2969

[13] Shah S A, Koltun V. Robust continuous clustering [J]. Proceedings of the National Academy of Sciences of the USA, 2017, 114 (37): 9814.

[14] Liu Guangcan, Lin Zhouchen, Yan Shuicheng, *et al.* Robust recovery of subspace structures by low-rank representation [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35 (1): 171-184.

[15] Lin Zhouchen, Liu Risheng, Su Zhixun. Linearized alternating direction method with adaptive penalty for low-rank representation [C]// Advances in Neural Information Processing Systems. 2011: 612-620.

[16] Lu Canyi, Min Hai, Zhao Zhongqiu, *et al.* Robust and efficient subspace segmentation via least squares regression [C]// Computer Vision. Berlin: Springer, 2012: 347–360.

[17] Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection [J]. Journal of the Royal Statistical Society, 2010, 72 (1): 3.

[18] Lu Canyi, Hai Min, Zhao Zhongqiu, *et al.* Robust and efficient subspace segmentation via least squares regression [C]. Berlin: Springer-Verlag, 2012: 347-360.

[19] Abdi H. Partial least squares regression and projection on latent structure regression (PLS Regression) [J]. Wiley Interdisciplinary Reviews Computational Statistics, 2010, 2 (1): 97–106.

[20] Lin Liyuan, Chen Xiaoyun, Jian C. Subspace segmentation via least squares regression including information about distance [J]. Microcomputer & Its Applications, 2016.

[21] Wang Shusen, Yuan Xiaotong, Yao Tiansheng, *et al.* Efficient subspace segmentation via quadratic programming [C]// Proc of the 25th AAAI Conference on Artificial Intelligence. 2011: 519-524.

[22] Shi Jianbo, Malik J. Normalized cuts and image segmentation [J]. IEEE Trans on Pattern Analysis & Machine Intelligence, 2000, 22 (8): 888-905.

[23] Xu Linli, Li Wenye, Schuurmans D. Fast normalized cut with linear constraints [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2009: 2866-2873.

[24] Elhamifar E, Vidal R. Sparse subspace clustering: algorithm, theory, and applications [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 35 (11): 2765-2781.

[25] Wang YuXiang, Xu Huan. Noisy sparse subspace clustering [J]. Journal of Machine Learning Research, 2013, 17 (1): I-89.

[26] Peng Xi, Zhang Lei, Yi Zhang. Scalable sparse subspace clustering [C]// Computer Vision and Pattern Recognition. 2013: 430-437.

[27] Zhu Xiaofeng, Zhang Shichao, Hu Rongyao, *et al.* Local and global structure preservation for robust unsupervised spectral feature selection [J]. IEEE Trans on Knowledge and Data Engineering, 2018, 30 (3): 517-529.